Volume 3, Issue 1

August 1994

National Institutes of Health

Mosaic Server Adds dbEST and dbSTS

The EST and STS databases can now be accessed through the NCBI's World Wide Web/Mosaic server, in the What's New section of the Home Page. The URL is http://www.ncbi.nlm.nih.gov. dbEST and dbSTS Mosaic pages offer information on the current release, text string retrieval of full reports, and information on data retrieval methods and data submission procedures. The Mosaic server also offers convenient searching of GenBank and Entrez.

Both databases are still available for similarity searching through NCBI's BLAST e-mail server. Full reports on selected matches can be obtained through the RETRIEVE server (retrieve@ncbi.nlm.nih.gov). For instructions on using the RETRIEVE server for dbEST or dbSTS, send the following help message to the address retrieve@ncbi.nlm.nih.gov:

datalib dbest (or dbsts) help

Note that this is a new address for obtaining full reports from dbEST. The est_report@ncbi.nlm.nih.gov server is no longer in use.

dbEST Update

Now in its second year of operation, dbEST is a resource that contains data from laboratories producing partial, single-pass cDNA sequences (ESTs or "Expressed Sequence Tags"). dbEST currently contains more than 50,200 cDNA sequences from 22 organisms as shown below.

dbEST Sequences

<i>Homo sapiens</i> 22,841
Caenorhabditis elegans 11,587
Arabidopsis thaliana 8,010
<i>Oryza sativa</i> 4,342
Plasmodium falciparum 1,104
<i>Zea mays</i>
Mus musculus+domesticus 393
Schistosoma mansoni 323
Brassica campestris 181
Saccharomyces cerevisiae 134
<i>Capra hircus</i> 108
Gallus gallus+domesticus 55
Pyrococcus furiosus 50
Brassica napus39
Macropus eugenii
Hippoglossus hippoglossus 32
Orcorhynchus mykiss22
Euplotes crassus
Pleuronectes americanus 14
Nicotiana tabacum13
Cyclopterus lumpus 12
Sus scrofa11
J

EST sequences are included in the EST division of GenBank, available from NCBI on CD-ROM and by Anonymous FTP. The sequences are also available as a flat file in the FASTA format by Anonymous FTP in the /repository/dbEST directory at ncbi.nlm.nih.gov.

dbSTS Update

First announced in the February 1994 issue of *NCBI News*, dbSTS facilitates access to the growing body of sequence tagged site (STS) data. Beginning with Release 85.0 of GenBank in October 1994, STS data

Continued on page 3

Entrez CD-ROM Begins Third Year With Third Disc

In October 1994, Entrez begins its third year of providing CD-ROM access to DNA and protein sequences and related bibliographic information. In the past 2 years, the number of subscribers has increased to more than 2,400 and the database size has more than doubled. To accommodate this rapid growth, Entrez added a second disc in February 1993 and will add a third disc this coming October.

New Year, New Disc, New Price

The addition of a third disc is factored into the new price of *Entrez*, as determined by the U.S. Government Printing Office (GPO). Effective in late September, the price of a 1-year subscription to *Entrez* will be \$102 in the United States, up from the current price of \$76. Current

Continued on page 4

In This Issue

Mosaic Adds dbEST/dbSTS 1
Entrez CD-ROM's Third Year 1
Repository CD-ROM Service Ends 2
GenBank Fellows 3
Entrez Replaces NCBI-Sequences 3
NCBI Data by FTP 3
Recent Publications 4
GenBank Flat File Expands 4
Profile: Dan Graur 5
Frequently Asked Questions 6
Special: GenBank Services

NCBI News

NCBI News is distributed three times a year. We welcome communication from users of NCBI databases and software and invite suggestions for articles in future issues. Send correspondence and suggestions to NCBI News at the address below.

NCBI News National Library of Medicine Bldg. 38A, Room 8N-803 8600 Rockville Pike Bethesda, MD 20894 Phone: (301) 496-2475 Fax: (301) 480-9241

E-mail: info@ncbi.nlm.nih.gov

Editors
Dennis Benson
Barbara Rapp

Design Consultant Troy M. Hill

Photography Karlton Jackson

Editing, Graphics, and Production Veronica Johnson Wendy B. Osborne

In 1988, Congress established the National Center for Biotechnology Information as part of the National Library of Medicine; its charge is to create automated systems for storing molecular biology, biochemistry, and genetics data, and to perform research in computational molecular biology.

The contents of this newsletter may be reprinted without permission. The mention of trade names, commercial products, or organizations does not imply endorsement by NCBI, NIH, or the U.S. Government.

NIH Publication No. 94-3272

ISSN 1060-8788

2

Repository CD-ROM Service Ends, FTP Access Continues

Since April 1992, the NCBI has distributed three complimentary releases of the Data Repository CD-ROM, which contains the same set of molecular biology databases available on the public FTP repository. Copies of the January 1994 edition are still available on request from NCBI, but no additional CD-ROMs will be produced. The FTP distribution site will continue to be maintained for access over the Internet by Anonymous FTP (see box on page 3).

The NCBI Data Repository was established to provide a public distribution site for contributed molecular biology databases. The table below lists the databases, abbreviated contents, and curators. Database contributors are solely responsible for the maintenance and accuracy of the collection. Information on how to contact curators is included in the README file for each database.

Questions, comments, and proposals for additions to the FTP repository should be sent to Scott Federhen at repository@ncbi.nlm.nih.gov.

Contents of NCBI Data Repository CD-ROM: January 1994				
Database	Content	Curator(s)		
aatdb	Arabidopsis thaliana	J. Morris		
acedb	Caenorhabditis elegans	R. Durbin/J. Thierry-Mieg		
aids-db	HIV and SIV sequences	G. Myers		
blocks	Protein homology blocks	S. Henikoff/J. Henikoff		
carbbank	Carbohydrate structure	D. Smith/S. Doubet/ P. Albersheim		
chrominfo	Human chromosome maps	P. Nadkami		
eco2dbase	Escherichia coli	F.C. Neidhardt		
enzyme	Enzyme information	A. Bairoch		
epd	Eukaryotic promoters	P. Bucher		
flybase	Drosophila	M. Ashburner		
genedict	Mouse genome information	R. Williams		
journals_toc	Tables of contents from selected journals	A. Bairoch		
kabat	Immunological sequences	E. Kabat		
limb	Molecular biology database directory	G. Redgrave		
metabolism	Intermediate metabolism compounds	P. Karp		
metproto	Metabolic reactions	R. Ochs		
ngdd	Normalized gene maps	Y. Abel/R. Cedergren		
nrfes	Nucleic acid sequences	A. Konopka		
pkinases	Protein kinase sequences	A.M. Quinn		
prosite	Protein sequence motifs	A. Bairoch		
rebase	Restriction enzymes	R. Roberts		
repbase	Human repetitive DNA sequences	J. Jurka		
rldb	DNA probe locations	G. Zehetner		
seqanalref	Sequence analysis bibliographies	A. Bairoch		
t4phage	T4phage	E. Kutter		
tfd	Transcription factors	D. Ghosh		
yeast	Yeast information	F. Ouellette		

GenBank Fellows To Focus Research on Database Quality

The GenBank Postdoctoral **▲** Fellowship Program is a new NCBI initiative to improve the quality of the GenBank database and also to serve as a bioinformatics training program. GenBank Fellows are selected for strong backgrounds in biology and the desire to apply computational tools to the organization of electronic data in molecular and structural biology, genetics, and phylogeny. NCBI provides training in the Unix operating system, software tools for manipulating data, sequence analysis methods and statistics, and database management systems.

Under the supervision of a mentor from NCBI's Computational Biology Branch, GenBank Fellows will pursue various applied research projects to improve the quality and annotation of GenBank entries, reduce sequence redundancy, and establish and maintain links to other databases such as those containing genetic and physical mapping data and three-dimensional macromolecular structures.

Four of the initial five positions have been filled from a pool of more than 200 applicants. Appointed for a minimum of 2 years, the first Gen-Bank Fellows began the program in July. For more information, contact David Landsman at NCBI. ■



Mosaic, continued from page 1

will be consolidated into a new division, called STS. The database currently contains nearly 6,000 DNA sequences from four organisms: *Homo sapiens* (4,763), *Drosophila melanogaster* (736), *Oryza sativa* (315), and *Bos taurus* (4).

EST and STS Data Submissions

EST and STS data may be submitted to GenBank via e-mail to the address: batch-sub@ncbi.nlm.nih.gov. GenBank accession numbers will then be assigned. For large-volume submissions, a temporary FTP account can be set up on request, in lieu of e-mail. NCBI staff welcomes test submissions for review of format prior to submission of an entire library. To obtain a copy of the data input format specification for EST or STS data or for further information, contact GenBank User Services at info@ncbi.nlm.nih.gov or (301) 496-2475. ■

Entrez Replaces NCBI-Sequences CD-ROM

The NCBI-Sequences (ASN.1) CD-ROM will be discontinued effective June 1995. Because *Entrez* includes the same data, plus additional indexes and a molecular sequence-related subset of MEDLINE, there is no longer a need for a separate CD-ROM. New subscribers should order *Entrez* instead.

If you have any questions, or for some reason require a single issue of the NCBI-Sequences CD-ROM rather than *Entrez*, contact GenBank User Services at NCBI. ■

NCBI Data by FTP

NCBI maintains a repository of molecular biology databases and software development tools that are publicly available for network users through Internet FTP (file transfer protocol). The available directories include "repository", "toolbox", and "pub".

The repository directory holds 27 databases, such as:

- Swiss-Prot (Amos Bairoch)
- ACeDB: A *C. elegans* Database (J. Thierry-Mieg, R. Durbin)
- FlyBase (Michael Ashburner)
- Eukaryotic Promoter Database (Philipp Bucher)
- REBASE (Restriction Enzyme Database - Richard Roberts)
- · CarbBank/CCSD
- PROSITE (Dictionary of Protein Sites and Patterns -Amos Bairoch)

The toolbox directory contains a set of software and data exchange specifications that are used by NCBI to produce portable software, and includes ASN.1 tools and specifications for molecular sequence data.

The pub directory offers public-domain software, such as BLAST (sequence similarity search program), MACAW (multiple sequence alignment program), and Authorin submission software for Mac and PC systems. Client software for Network *Entrez* and Network BLAST is also included in this directory.

Data in these directories can be transferred through the Internet by using the Anonymous FTP program. To connect, type: ftp ncbi.nlm.nih.gov or ftp 130.14.25.1. Enter anonymous for the login name, and enter your email address as the password. Change directories to "repository" to download databases (cd repository), "toolbox" to download ASN.1 tools (cd toolbox), or "pub" to download public-domain software (cd pub).



Selected Recent Publications by NCBI Staff

Altschul, SF, MS Boguski, W Gish, and **JC Wootton.** Issues in searching molecular sequence databases. *Nature Genet* 6:119–29, 1994.

Boguski, MS. Bioinformatics. *Curr Opin Genet Dev* 4:383–8, 1994.

Hartmann, C, **H Recipon**, MF Jubier, C Valon, E Delcher-Bessin, Y Henri, J DeBuyser, B Lejeune, and A Rode. Mitochondrial DNA variability in a single wheat regenerant involves rare recombination event across a short repeat. *Curr Genet* 25: 456–64, 1994.

Kokubo, T, DW Gong, **JC Wootton**, M Horikoshi, RG Roeder, and Y Nakatani. Molecular cloning of *Drosophila* TFIID subunits. *Nature* 367:484–7, 1994.

Rapp, BA. Information Resources in Biotechnology: North and South America. (Medford, NJ: Learned Information, Inc.), 1994.

Senkevich, TG, **EV Koonin**, and RML Buller. A poxvirus protein with a RING finger motif is of crucial importance for virulence. *Virology* 198:187–97, 1994.

Spouge, JL. Viral multiplicity of attachment and its implications for HIV therapies. *J Virol* 68:1782–9, 1993.

States, DG, and **W Gish**. Combined use of sequence similarity and codon bias for coding region identification. *J Comput Biol* 1:39–50, 1994.

Tugendreich, S, **MS Boguski**, MS Seldin, and P Hieter. Linking yeast genetics to mammalian genomes: Identification and mapping of the human homolog of *CDC27* via the expressed sequence tag (EST) data bases. *Proc Natl Acad Sci U S A* 90:10031–5, 1993.

Wilbur, WJ, and L Coffee. The effectiveness of document neighboring in search enhancement. *Inf Process Manage* 30:253–66, 1994.



GenBank Flat-File Expands to Two CD-ROMs

Beginning with Release 84.0 to be distributed in late August, subscribers to the GenBank Flat-File CD-ROM will receive the database on two CD-ROMs. The increase to two discs is necessary due to the steady increase in the size of GenBank. Release 84.0 will contain more than 201,815,802 nucleotides from more than 196,703 reported sequences, and will require 695 MB to store the database and accompanying index files. A price increase to \$66 will go into effect in October.

The GenBank Flat-File CD-ROM contains the GenBank database only, and includes no software. Because this version of GenBank is primarily used for data distribution, with the data being transferred to magnetic disk for use with locally implemented sequence search and analysis software, this increase to two discs should not pose any difficulties. Simply copy the contents of both CD-ROMs to your magnetic disk. If you encounter any problems in using the two-disc set, contact NCBI. ■



Entrez CD-ROM, continued from page 1

domestic and foreign prices will be honored for orders placed before the publication of the new prices.

Subscription Renewals

The GPO mails subscription renewal notices approximately 90 days prior to the end of annual subscriptions. Subscriptions are not automatically renewed, so be sure to reorder early enough to avoid an interruption in service.

What Do I Do With Three Discs?

You have several options to accommodate the additional *Entrez* disc, as reported in the February issue of *NCBI News*. A new option requires no changes at all. In response to concerns expressed by some subscribers, we have designed the three-disc *Entrez* so that it will be possible to use *Entrez* on PCs and Macs with a single CD-ROM drive, although you will need to swap discs whenever your query or lookup requires a different database. Using the following options will yield higher performance:

- 1. Install a total of three CD-ROM drives. On a Mac, it is as simple as making sure the SCSI IDs are unique and plugging in the drives. On a PC, make sure you have SCSI CD-ROM drives and that the "device driver" software supports multiple drives. With "double-speed" SCSI CD-ROM drive prices now about \$300, this will be your lowest cost solution for high performance.
- 2. Use the *Entrez* CD-ROMs only as a distribution medium and transfer all the files and data to magnetic disk. This option will require about 2 gigabytes of disk space. Although more expensive, this approach yields the highest performance. One hard disk-based "server" could provide *Entrez* access for an entire department. The 2 gigabytes of additional storage should cost less than \$2,000.
- 3. The last alternative, Network *Entrez*, is free if you already have a direct TCP/IP connection to the Internet. Use Anonymous FTP to obtain the Internet version of *Entrez*

Continued on page 6

PROFILE

Dan Graur: Classical Challenger

Is the guinea pig a rodent? In a 1991 article by this title (*Nature* 351:649-52), Dan Graur challenged classical taxonomy by showing that the rodents are not evolutionarily monophyletic. He demonstrated that the myomorph rodents (e.g., rats, mice, and squirrels) are closer to primates and artiodactyls than they are to the hystricomorphs (e.g., guinea pigs and porcupines), thereby rendering the taxon Rodentia invalid. Then in 1994, he further "deconstructed" classical taxonomy by showing that the cetaceans (whales and dolphins) are in fact more closely related to the ruminants than either taxa is to other even-hoofed ungulates, such as pigs and camels (Mol Biol Evol 11:357-64).

A leading expert in molecular evolution and phylogeny, Graur is currently on sabbatical from his position as Associate Professor at the Department of Zoology, Tel Aviv University. At NCBI, his goal has been to use molecular sequence data to reconstruct the phylogeny of all mammalian orders.

New Protocol

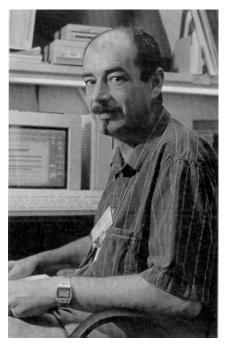
Graur explains that there are several tradeoffs in reconstructing higher level phylogenetic trees based on molecular data. For example, the size of taxonomic sample is inversely related to the number of sequences that can be used in an analysis. He has therefore developed a protocol to maximize the amount of molecular data that can be used in a phylogenetic study by producing sets of orthologous sequences, each representing a monophyletic taxon. For each such set, the number of amino

acid replacements is inferred, and the results are tabulated according to the number of informative sites that support a certain clustering of two taxa to the exclusion of a third. The entire procedure is repeated for each group of orthologous proteins separately. Graur notes that it is very slow and tedious work, but he finds the results satisfying.

Point of Order

Preliminary results indicate that the available data may already be sufficient to elucidate the ordinal phylogeny of the eutherian mammals—or at least to obtain a fairly accurate outline of the phylogenetic relationships. By this, Graur means that at the end of the study, the phylogenetic tree of the eutherian mammals will either be solved in a binarily satisfying manner, or will be reduced to only a few alternative trees (out of the approximately 10^{26} possible trees) for further consideration.

Graur says the emerging taxonomic pattern is interesting because it seems to invalidate all the superordinal clusterings thus far proposed in the morphological and paleontological literature. Many superorders, such as Glires (rabbits and rodents), Anagalidia (elephant shrews and rodents), and Ungulata (perissodactyls, artiodactyls, hyracoids, and proboscids), have been rendered invalid, and those that remain will need to be redefined. The study also sheds light on the earliest divergence events in the evolution of placental mammals. It seems, for example, that contrary to conventional morphological belief, the insectivores rather than the



Dan Graur, on sabbatical at NCBI, conducts research in molecular engineering and phylogeny.

Edentata seem to represent the earliest branch within placental evolutionary history.

The reaction to this painstaking effort at evolutionary elucidation? Morphologists and paleontologists are not amused. One of Graur's colleagues observed: "If guinea pigs are not rodents, if whales are cows, if rabbits are primates, and carnivores are related to horses, then maybe molecular biology is for the birds."

Dan Graur received his B.Sc. in Biology and M.Sc. in Zoology from Tel Aviv University, and his Ph.D. in Genetics from the University of Texas. An Associate Professor in the Department of Zoology at Tel Aviv University, he also serves as editorin-chief of the Israel Journal of Zoology and is a member of the editorial board of Molecular Phylogenetics and Evolution.

NCBI News • August 1994 5



Frequently Asked Questions

When formatting queries for the RETRIEVE server, do Ineed to worry about upper and lower case?

The RETRIEVE server is case-insensitive so it doesn't matter if any part of the query is in capital or small letters.

Why does the record with the same accession number sometimes appear in GenBank and GenBank Updates?

When a record is revised, the updated version goes into the GenBank Updates and the original version remains in the current full release. With a new release, the updated version replaces the original version.

What is the significance of the NCBI gi number that appears in a GenBank record?

NCBI gi numbers are assigned to every nucleotide and protein sequence loaded into the GenBank database. This number supplements the accession number. The gi is used to track updates in the sequence information. Any change in a sequence results in the generation of a new gi number, even though the accession number remains the same.

If I have a subscription to Entrez on CD-ROM, should I also get Network Entrez?

Network *Entrez* is available free of charge, but you must have direct TCP/IP access to the Internet and a local systems administrator to install the software and provide end-user support. The sequence information on the Network and CD-ROM versions of *Entrez* is identical and is updated at the same time. With Network *Entrez*, there is an option to access a larger subset of MEDLINE.

Why are recently published accession numbers sometimes not available through RETRIEVE and BLAST?

Occasionally when authors request that their information be held until publication, they do not notify us of publication. To have data released, send a complete bibliographic citation—journal name, article title, author(s), volume, page number, and year—along with the accession number to update@ncbi.nlm.nih.gov. GenBank staff will verify publication and release the record.

Entrez CD-ROM, continued from page 4

specific to your computer (Windows PC, Macintosh, and several different Unix systems). The software may be found in the /entrez/network directory of the NCBI's FTP site (ncbi.nlm.nih.gov). It will be necessary to register your subnet address with the NCBI. Your local

network administrator is responsible for maintaining the TCP/IP network software, a functioning Internet connection, and providing end-user support. For more information, see the README file or contact net-info@ncbi.nlm.nih.gov.

GenBank: Easy Deposits, Unlimited Withdrawals, High Interest

It's easy—and free—to contribute sequences to GenBank and search the database. This table summarizes the data submission and search services available from NCBI.

Service	Purpose	How To Use/How To Get Help
GenBank submissions	Submit new sequences to GenBank.	Send a new submission by e-mail: gb-sub@ncbi.nlm.nih.gov
GenBank updates	Correct or update an existing sequence; request release of published data.	Send an update request by e-mail: update@ncbi.nlm.nih.gov
Authorin submission software	Prepare new or updated GenBank data.	Request Authorin software or help with a new submission or update: authorin@ncbi.nlm.nih.gov or call the Service Desk at (301) 496-2475.
Automated e-mail services		
retrieve@ncbi.nlm.nih.gov	Retrieve GenBank and other sequence database records from an e-mail server based on any text term, including accession number, author name, locus or gene name,	To receive documentation, send a message containing only the word "help" (unquoted), in the body of the message.
	etc.	To receive personal assistance, send e-mail to: retrieve-help@ncbi.nlm.nih.gov
blast@ncbi.nlm.nih.gov	Perform a sequence similarity search of GenBank and other sequence databases using the BLAST algorithm.	To receive documentation, send a message containing only the word "help" (unquoted) in the body of the message. To receive personal assistance, send e-mail to: blast-help@ncbi.nlm.nih.gov
Network applications	"Client-server" programs, in which client program on local PC, Mac, or Unix workstation queries NCBI server via the network.	All NCBI network applications require Internet access and locally installed TCP/IP software.
Network Entrez	Point-and-click retrieval system for Windows PCs, Macs, and Unix workstations. For text-based searching of sequence databases and a sequence-related subset of MEDLINE.	For registering and obtaining client software, send e-mail to: net-info@ncbi.nlm.nih.gov
Network BLAST	Interactive BLAST similarity searching for Windows PCs, Macs, and Unix workstations.	For registering and obtaining client software, send e-mail to: blast-help@ncbi.nlm.nih.gov
Mosaic/World Wide Web NCBI Home Page URL: http://www.ncbi.nlm.nih.gov	Hypertext-like access to NCBI databases and search services, including Mosaic versions of RETRIEVE, BLAST, and <i>Entrez</i> .	Requires access to Internet "browsing" software, such as Mosaic. Ask your local network administrator about WWW/ Mosaic.
Anonymous FTP: ncbi.nlm.nih.gov	Obtain GenBank releases, NCBI software, and various molecular biology databases.	Login as "anonymous" (unquoted) and enter your e-mail address as your password.
CD-ROMs	For users who do not have Internet access or who prefer local copy of databases.	For information about subscriptions, send e-mail to: info@ncbi.nlm.nih.gov
Entrez (GPO list ID: ENT)	CD-ROM version of Network <i>Entrez</i> . One year subscription (6 issues, 3 discs/issue): \$102.	Send e-mail questions to: entrez@ncbi.nlm.nih.gov
GenBank (GPO list ID: NCBIF)	GenBank in "flat-file" format, as required by some commercial and academic software. One year subscription (6 issues, 2 discs/issue): \$66.	Send e-mail to: info@ncbi.nlm.nih.gov

To Receive Information From The NCBI, Please Send This Form To:			
NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION National Library of Medicine National Institutes of Health Bldg. 38A, Room 8N-803 8600 Rockville Pike Bethesda, MD 20894	Queries about services and software may also be sent via electronic mail to info@ncbi.nlm.nih.gov or by fax to (301) 480-9241.		
Name (please print) Department University or Institution Mailing Address	NCBI News Mailing List (Check one) Add new name to NCBI News mailing list Remove name from NCBI News mailing list		
Phone Fax E-mail	Make corrections or changes as shown (please enclose label)		
Comments			

DEPARTMENT OF HEALTH AND HUMAN SERVICES Public Health Service, National Institutes of Health

National Library of Medicine National Center for Biotechnology Information Bldg. 38A, Room 8N-803

8600 Rockville Pike Bethesda, Maryland 20894

FIRST-CLASS MAIL POSTAGE & FEES PAID PHS/NIH/NLM BETHESDA, MD PERMIT NO. G-763

Official Business Penalty for Private Use \$300